Enzyme family coherence assessment: validation and prediction

H. P. Bastos,[1] T. Grego,[1] F. M. Couto,[1] and P. M. Coutinho[2]

[1] *Faculdade de Ciências da Universidade de Lisboa*
[2] *Architecture et Fonction des Macromolécules Biologiques, UMR6098, CNRS, Universités Aix-Marseille*

MOTIVATION:

The sequencing of multiple genomes is currently progressing at a fast pace. Each protein sequence needs to be functionally annotated in order to acquire a biological context. The high quality manual annotations made by specialized curators, is unable to keep up with the rhythm at which new organisms are being sequenced, which has led to the development of automatic annotation methods. However, these methods are based on existing annotations, and thus biased toward the best studied and characterized model organisms. Furthermore, existing annotation errors are further propagated by these methods. This justifies improvement of the automatic methods so that their precision is increased to a more acceptable level. A current compromise resides in the use of semi-automatic methods, where the automatic procedures are employed to propose annotations that require validation by specialized curators. However, any methods capable of reducing human intervention (up to full automation) without loss of precision are certainly very desirable.

BACKGROUND:

The developments of automatic methods for protein functional annotation aims to accompany the genome sequencing rhythm. As an example, Swissprot database, which contain only manually annotated sequences, contains 18 times less sequences than TrEMBL, which houses 6.6 millions sequences, being that the latter are automatically analysed and annotated [1].

Most of the existing methods for automatic protein functional annotation share a first common step, the identification of homologue sequences, that is, identification of sequences with a common evolutionary origin [2]. A simple example that uses this approach is GeneQuiz [3] which annotates proteins by using only the information of the most informative sequence selected from a set of homologue sequences.

However the use of homologue sequences does not guarantee correct automatic functional annotation [4] since it generates systematic errors, that are derived from the following biological phenomena: gene duplication, evolutionary distance and domain shuffling [5]. The first two explain why reasonably similar proteins can have different functions, and the domain shuffling reveals the dynamic of the genome, confirming that homology regions can be only local.

In order to circumvent these issues, several approaches were developed using not only one sequence but information from multiple homologue proteins and their respective alignments, and annotation sharing [6–8]. However the propagation of undetected misannotations in databases is still an unsolved issue [5]. It was estimated that even in curated processes that use sequence similarity methodology the annotation error rate can reach 49% [9]. Several protein databases, such as SYSTERS [10] and ClustR [11] use primarily this type of strategy in their automatic annotation procedures. However, to achieve a compromise between annotation coverage and precision, semi-automatic processes such as the ones employed on COG [12], CATH [13] or CAZy[14] exist. These methods typically use automatic procedures that require verification by specialists. Hence, they are mostly used on specialized databases.

Other automatic methods were developed to face the issue caused by the growing number of misannotations, these methods try to correct or validate annotations [15, 16]. One of such

methods is CAC [15] that validates predicted annotations by correlating with previously manually curated annotations. MisPred [16] also does this validation but using a set of rules based on previous observations of biological knowledge. Beyond annotation validation, the recent use of ontologies in the annotation process has allowed uniformity and ambiguity removal from protein annotation [17]. Gene Ontology (GO), is nowadays a standard in the community and provides a controlled vocabulary to describe genes and attributes of gene products of any organism. Gene Ontology is composed by three different ontologies that describe the gene products in terms of biological process, molecular function and cellular component [18].

OBJECTIVE:

With this work we aim to develop computational methods that calculate the robustness of functional annotation of protein families. These measures will enable us to identify incoherently annotated groups of sequences in order to make them the target of intervention by expert curators and/or experimental biochemical characterization. Under-annotated families will be complemented with annotation identified by new methods of text mining, which will also be developed. Families that are identified as robust will serve as knowledgebase to generate an ontology extension for submission to Gene Ontology. With this submission of new terms to the Gene Ontology we intend to enrich it in the area of Glycobiology, because there are still not enough GO terms available neither in quantity or specificity to cover most of the biological processes of this area.

As a case-study we will use the resources of CAZy (`www.cazy.org`), a source of knowledge in the area of Glycobiology, which is one of the main databases specialized in families of structurally-related catalytic and carbohydrate-binding modules (or functional domains) of enzymes that degrade, modify, or create glycosidic bonds [14].

APPROACH:

We started by performing an exploratory analysis of CAZy public data. Currently it contains over 140,000 proteins obtained from public databases and through collaborations with international consortia of genomic annotation, including about 7,000 enzymatic activities and adhesion extracted from over 30,000 bibliographical references. It is organized into 290 protein families covering five classes of enzymatic activities: Glycoside hydrolases (GH), glycosyltranferases (GT), polysaccaride lyases (PL), carbohydrate esterases (CE) and carbohydrate-binding modules families (CBM) [14]. To assess the functional coherence of the CAZy families we measured the intra-family semantic similarity. We used the simGIC semantic similarity measure [19] on the Uniprot entries of the CAZy families. We used only the UniProt entries since only these were directly linked to GO term annotations. In fact, we have seen that on average only 77% of uniprot entries in each CAZy family had GO term annotations from the *molecular_function* ontology. We have plotted (Figure 1) the distribution of sizes (in number of uniprot entries) of the CAZy families with less than 1000 entries (only 15 families had more than 1000) and we randomly sampled families to represent the five different enzymatic activities classes at five peaks in the graphic. Their *molecular_function* ontology coverage over the number of Uniprot entries per family and over the total number of all entries per family can be seen in Table 1.

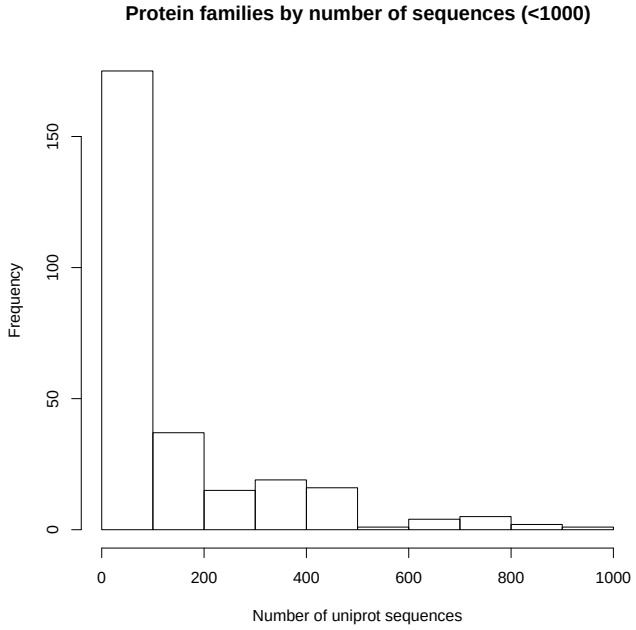

FIG. 1: CAZy family size distibution. Each bar shows the number of families (frequency) with a given number of sequences.

Family	uniprot coverage	%GO annot. over uniprot	%GO annot. over total	Family	uniprot coverage	%GO annot. over uniprot	%GO annot. over total
CE12	53	100	31	PL1	331	78	7
PL9	55	60	23	GH7	351	100	38
CBM9	55	100	41	GT25	359	58	23
GT56	57	98	34	CE10	378	86	25
GH84	60	40	12	CE9	576	98	37
PL3	136	99	37	GT35	701	99	33
CE5	150	99	31	GH19	744	94	31
GH92	150	27	8	CE1	944	51	18
CBM3	151	100	35	GT9	1042	98	40
GT6	161	100	25	GH1	1146	99	29
CBM50	347	52	20	CBM48	1392	96	33

Table 1: Number of Uniprot entries per family, percentage of Uniprot entries in a family annotated with GO *molecular_function* terms, and ratio of GO *molecular_function* annotated terms per total number of family entries.

RESULTS AND DISCUSSION:

The semantic similarity profiles of the sampled CAZy families gave us insight into their coherence. As expected, all the sampled CBM families (four) showed us varying degrees of semantic similarity (CBM3 family shown Figure 2a)). This is not surprising since these families comprise of members that often associate themselves to other carbo-active catalytic modules in the same polypeptide and can target different substract forms depending on different structural characteristics [14]. The most coherent families were shown to be CE12,

GT56, PL3 and GH7 (GT56 and PL3 shown at Figure 2b) and c) respectively), being that GT56 scored a perfect semantic similarity of 1 for all its pairs of (Uniprot) proteins. This happens because all of the considered proteins from this family were annotated to the same high informative term, *fucosyltransferase activity*. The family PL3 yielded similar results (Figure 2c)) since most of its proteins were annotated with the term *pectate lyase activity*. However most of the sampled cases we observed (GH92, GT6, CE5, GT25, PL1, CE9, GT35, GH19, GT9 and GH1) showed a configuration where two peeks of similarity arose, one at the far right of the histogram and another one before the 0.5 semantic similarity threshold. We can see two of these cases at Figure 2d) for family GT9 and Figure 2e) for family CE5. What happens here is that in the case of family GT9, most of the proteins are annotated with the term *transferase activity* while only about half of those are also annotated with a more specific term *transferase activity, transferring glycosyl groups*. Again the same behaviour is observed in family CE5, with most of the terms being annotated with the term *hydrolase activity*, and only half of them having also a more specific *cutinase activity* term annotated to them.

Family GH84 (Figure 2f)) as can be seen on Table 1 is an example of where the GO ontology still offers poor coverage. Only 40% of the Uniprot entries of this family have GO *molecular_function* annotations and that covers only 12% of all the entries in the family. Although there are only 24 proteins in this family that contain GO *molecular_function* annotations, the specificity up to each protein is annotated varies greatly. Hence, on Figure 2f) we can see peeks at five different similarity levels. This, however, does not mean lack of family coherence but instead it means there is an uneven depth of GO annotation. Even though GO annotations still lack in specificity, on average they cover 77% of the Uniprot entries in each family. However, if we consider consider all the entries in each family, and not just those from Uniprot, our sample data presents a coverage from 7% to 41%, which gives an average of only 28%. This means there is still a big percentage of data to explore. Hence in the future we will have to employ ways to extract useful information from the other non-Uniprot entries. This can be combined with the semantic similarity measures on a method for the complete assessment of coherence for every CAZy family.

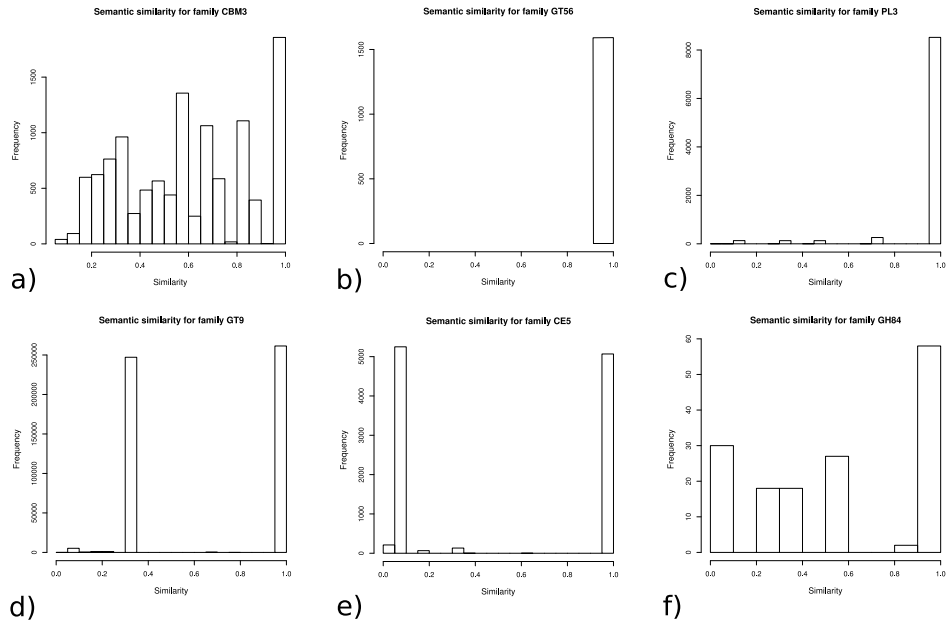

FIG. 2: Frequency distributions of semantic similarities between pairs of proteins in a CAZy family. Plots for family a) CBM3, b) GT56, c) PL3, d) GT9, e) CE5 and f) GH84.

CONCLUSIONS:

So far we have successfully used semantic similarity to identify some coherent protein families. In the process, we were also able to identify shortcomings in the depth of annotation of GO annotations for some of the Uniprot proteins in the CAZy families. There is still, however, more information from over 70% of family entries that can be potentially used to perform robust coherence assessments. Thus, by now, we have can discover which are the CAZy families with functional annotation robustness and that enable us to select some candidate families with members that require deeper ontology annotation. Soon, we will be applying this method to the complete CAZy family space leading us to the refinement of this process. We will also start to harvest and assess how to use the information provided by non-Uniprot entries, and apply text mining techniques, when required, so that all of these tasks can be effectively and automatically performed over a larger family space.

[1] The-Uniprot-Consortium, Nucleic Acids Res **36** (2008), ISSN 1362-4962.

[2] L. D. Stein, Nature Reviews Genetics **2**, 493 (2001), ISSN 1471-0056.

[3] M. Andrade, N. Brown, C. Leroy, S. Hoersch, A. de Daruvar, C. Reich, A. Franchini, J. Tamames, A. Valencia, C. Ouzounis, et al., Bioinformatics **15**, 391 (1999).

[4] P. Bork and E. Koonin, Nat Genet **18**, 313 (1998).

[5] D. Brown and K. Sjlander, PLoS Comput Biol **2**, e77 (2006).

[6] M. A. Andrade, in *Proceedings of the Seventh International Conference on Intelligent Systems for Molecular Biology* (AAAI Press, 1999), pp. 28–43, ISBN 1-57735-083-9.

[7] W. Fleischmann, S. Mller, A. Gateau, and R. Apweiler, Bioinformatics **15**, 228 (1999).

[8] F. Abascal and A. Valencia, Proteins **53**, 683 (2003).

[9] C. Jones, A. Brown, and U. Baumann, BMC Bioinformatics **8**, 170 (2007).

[10] A. Krause, J. Stoye, and M. Vingron, BMC Bioinformatics **6**, 15 (2005).

[11] R. Petryszak, E. Kretschmann, D. Wieser, and R. Apweiler, Bioinformatics **21**, 3604 (2005).

[12] R. Tatusov, M. Galperin, D. Natale, and E. Koonin, Nucleic Acids Res **28**, 33 (2000).

[13] F. M. G. Pearl, C. F. Bennett, J. E. Bray, A. P. Harrison, N. Martin, A. Shepherd, I. Sillitoe, J. Thornton, and C. A. Orengo, Nucl. Acids Res. **31**, 452 (2003).

[14] B. L. Cantarel, P. M. Coutinho, C. Rancurel, T. Bernard, V. Lombard, and B. Henrissat, Nucl. Acids Res. **37**, D233 (2009).

[15] F. M. Couto, M. J. Silva, and P. M. Coutinho, in *CIKM '06: Proceedings of the 15th ACM international conference on Information and knowledge management* (ACM, New York, NY, USA, 2006), pp. 142–151, ISBN 1-59593-433-2.

[16] A. Nagy, H. Hegyi, K. Farkas, H. Tordai, E. Kozma, L. Banyai, and L. Patthy, BMC Bioinformatics **9**, 353 (2008).

[17] H. Xie, A. Wasserman, Z. Levine, A. Novik, V. Grebinskiy, A. Shoshan, and L. Mintz, Genome Res. **12**, 785 (2002).

[18] M. Ashburner, C. Ball, J. Blake, D. Botstein, H. Butler, J. Cherry, A. Davis, K. Dolinski, S. Dwight, J. Eppig, et al., Nat Genet **25**, 25 (2000).

[19] C. Pesquita, D. Faria, H. Bastos, A. Ferreira, A. Falco, and F. Couto, BMC Bioinformatics **9**, S4 (2008).